SVM Speaker Verification using Session Variability Modelling and GMM Supervectors

M. McLaren, R. Vogt, S. Sridharan

Speech and Audio Research Laboratory
Queensland University of Technology, Brisbane, Australia
{m.mclaren, r.vogt, s.sridharan}@qut.edu.au

Abstract. This paper demonstrates that modelling session variability during GMM training can improve the performance of a GMM supervector SVM speaker verification system. Recently, a method of modelling session variability in GMM-UBM systems has led to significant improvements when the training and testing conditions are subject to session effects. In this work, session variability modelling is applied during the extraction of GMM supervectors prior to SVM speaker model training and classification. Experiments performed on the NIST 2005 corpus show major improvements over the baseline GMM supervector SVM system.

1 Introduction

Commonly, text-independent speaker verification systems employ Gaussian mixture models (GMMs) trained using maximum a-posteriori (MAP) adaptation from a universal background model (UBM) to provide state-of-the-art performance [1,2,3]. The GMM-UBM approach involves generative modelling whereby the distribution that produced some observed data is determined.

A major challenge in the design of speaker verification systems is the task of increasing robustness under adverse conditions. During the GMM training process, adverse effects due to session variability contribute to errors in the distribution estimation.

Recently, a method was proposed to directly model session variability in telephony speech during model training and testing [4,5]. The main assumption is that session effects can be represented as a set of offsets from the true speaker model means. A further assumption is that the offsets are constrained to a low-dimensional space. Session variability modelling attempts to directly model the session effects in the model space, removing the need for discrete session categories and data labelling required for regular handset, channel normalisation and feature mapping [6,7]. Direct modelling of session effects has led to a significant increase in robustness to channel and session variations in GMM-based speaker verification systems. Results show a reduction of 46% in EER and 42% in minimum detection cost over baseline GMM-UBM performance on the Mixer corpus of conversational telephony data [5].

Session variability modelling aims to relate the session effects across the mixture components of a model. For this technique, the speaker dependent information of a model's mixture components can be conveniently represented as a GMM mean supervector formed through the concatenation of the GMM component means.

In contrast to the traditional GMM-UBM classifier, the support vector machine (SVM) is a two-class, discriminative classifier whereby the maximum margin between classes is determined. SVMs utilise a kernel to linearly separate two classes in a high-dimensional space. A SVM speaker verification system recently presented by Campbell et al. has utilised GMM mean supervectors as features to provide performance comparable to state-of-the-art GMM-UBM classifiers [8].

The fundamental differences between GMM and SVM classification bring into question whether techniques used to improve GMM systems based on *distribution estimation* can also enhance SVM classification based on *margin maximisation*. This paper aims to demonstrate that robust modelling techniques developed for generative modelling can improve the performance of discriminative classifiers. The approach taken involves modelling session variability in the GMM mean supervector space prior to SVM speaker model training and classification.

A description of the common GMM-UBM system and recent research into session variability modelling is presented in Section 2. A brief summary of support vector machines is presented in Section 3 along with details regarding the extraction of session variability modelled supervectors for SVM training. Presented in Section 4 is the experimental configuration with the results of the system evaluated using the NIST 2005 database in Section 5.

2 Modelling Session Variability in the GMM Mean Supervector Space

2.1 The GMM-UBM classifier

In the context of speaker verification, GMMs are trained using features extracted from a speech sample to represent the speech production of a speaker. In such a system, MAP adaptation is employed to adapt only the means of the UBM to model a speaker [1]. The classifier computes a likelihood ratio using the trained speaker model and the UBM, giving a measure of confidence as to whether a particular utterance was produced by the given speaker. GMM-UBM classifiers provide state-of-the-art performance when coupled with a combination of robust feature modification and score normalisation techniques [3,9].

The GMM likelihood function is,

$$g(x) = \sum_{c=1}^{C} \boldsymbol{\omega}_c \mathcal{N}(\boldsymbol{x}; \boldsymbol{\mu}_c, \boldsymbol{\Sigma}_c), \tag{1}$$

where $\boldsymbol{\omega}_c$ are the component mixture weights, $\boldsymbol{\mu}_c$ the means, and $\boldsymbol{\Sigma}_c$ the covariances of the Gaussians. A mean supervector can be obtained by concatenating

each of the mean vectors, $\boldsymbol{\mu} = \left[\boldsymbol{\mu}_1^T \ \cdots \ \boldsymbol{\mu}_C^T\right]$. As only the means are adapted during speaker model training, the speaker model can be compactly represented by the common UBM and a speaker dependent GMM mean supervector offset.

2.2 Session Variability Modelling

Attempts to directly model session variability in GMM-UBM based speaker verification systems have provided significant performance improvements when using telephony speech [5]. The purpose of session variability modelling is to introduce a constrained offset of the speaker's mean vectors to represent the effects introduced by the session conditions. In other words, the Gaussian mixture model that best represents the acoustic observations of a particular recording is the combination of a session-independent speaker model and an additional session-dependent offset from the true model means. This can be represented in terms of the GMM component means supervectors as

$$\boldsymbol{\mu}_h(s) = \boldsymbol{m} + \boldsymbol{y}(s) + \boldsymbol{U}\boldsymbol{z}_h(s). \tag{2}$$

Here, the speaker s is represented by the offset $\boldsymbol{y}(s)$ from the speaker independent (or UBM) mean supervector $\boldsymbol{m}$. To represent the conditions of the particular recording (designated with the subscript h), an additional offset of $\boldsymbol{U}\boldsymbol{z}_h(s)$ is introduced where $\boldsymbol{z}_h(s)$ is a low-dimensional representation of the conditions in the recording and $\boldsymbol{U}$ is the low-rank transformation matrix from the constrained session variability subspace to the GMM mean supervector space.

Speaker models are trained through the simultaneous optimisation of the model parameters $\boldsymbol{y}(s)$ and $\boldsymbol{z}_h(s), h = 1, ..., H$ over a set of training observations. The speaker model parameters are optimised according to the *maximum a posteriori* (MAP) criterion often used in speaker verification systems [10,2]. The speaker offset $\boldsymbol{y}(s)$ has a prior as described by Reynolds [1] while the prior for each of the session factors $\boldsymbol{z}_h(s)$ is assumed to belong to a standard normal distribution, $\mathcal{N}(\boldsymbol{0}, \boldsymbol{I})$.

An efficient procedure for the optimisation of the model parameters is described in [5]. The session variability vectors are not actually retained to model the speaker but their estimation is necessary to accurately estimate the true speaker means. A similar optimisation process is used during testing.

3 Support Vector Machines

A support vector machine (SVM) performs classification by mapping observations to a high-dimensional, discriminative space while maintaining good generalisation characteristics [11]. SVM training involves the positioning of a hyperplane in the high-dimensional space such that the maximum margin exists between classes; a procedure unlike distribution estimation for GMMs. The term *support vectors* refers to the training vectors which are located on or between the class boundaries and, as a result, contribute to the positioning of the separating

hyperplane. A kernel function $K(\boldsymbol{X}_a, \boldsymbol{X}_b) = \phi(\boldsymbol{X}_a) \cdot \phi(\boldsymbol{X}_b)$ is used to compare observations in the high-dimensional space to avoid explicitly evaluating the mapping function $\phi(\boldsymbol{X})$.

3.1 A GMM Supervector SVM Speaker Verification System

In the context of speaker verification, it is necessary to be able to compare two utterances of varying lengths when using SVMs. A method of achieving this is to train a GMM through mean adaptation to represent each utterance, from which mean supervectors can be extracted. SVMs using GMM supervectors as feature vectors have demonstrated promising capabilities when only feature mapping and feature normalisation are applied [8].

The process of producing a GMM mean supervector to represent an utterance can be viewed as a kernel. Essentially, features from a variable length sequence of feature vectors are being transformed from the input space to the SVM *feature space*. In the given context, the SVM feature space has a dimension determined by the length of the GMM mean supervector.

The SVM system implemented in this work uses the mean offsets from a gender dependent UBM as input supervectors for SVM classification. That is, the supervector representing utterance $\boldsymbol{X}_a$ is the difference between the supervector $\boldsymbol{\mu}_a$ extracted from the mean adapted GMM trained from $\boldsymbol{X}_a$ and the supervector $\boldsymbol{m}$ taken from the gender dependent UBM. The motivation for removing the UBM mean bias is to reduce the rounding errors accumulated when equating dot products of floating point representations in high dimensions.

The input supervectors are also scaled to have unit variance in each dimension based on statistics from the background dataset. The aim of this process is to allow each dimension of the supervector an equal opportunity to contribute to the SVM.

The SVM kernel using background data scaling can then be formulated as,

$$K(\boldsymbol{X}_a, \boldsymbol{X}_b) = (\boldsymbol{\mu}_a - \boldsymbol{m})^T \boldsymbol{B}^{-1} (\boldsymbol{\mu}_b - \boldsymbol{m}), \tag{3}$$

where $\boldsymbol{B}$ is the diagonal covariance matrix of the background data. This background dataset is a collection of non-target speakers used to provide negative examples in the SVM training process.

3.2 Incorporating Session Variability into GMM Supervectors

The session variability modelling technique described in Section 2.2 is employed during GMM training to estimate and remove the contribution of the session conditions from the adapted model means. The trained model means can then be represented in GMM supervector form by $\boldsymbol{y}(s)$ in (2). This session-independent speaker model provides a method of incorporating session variability modelling into SVM classification. This differs from Campbell's nuisance attribute projection (NAP) in which subspaces in the SVM kernel contributing to variability are removed through projection [12].

The following experiments attempt to model session variability into the GMM supervectors during GMM training in order to demonstrate the possible advantages that such techniques for generative modelling may impart on discriminative classification.

4 Experiments

Evaluation of the proposed method was conducted using the NIST 2005 speaker recognition corpus consisting of conversational telephone speech from the Mixer Corpus. Focus was given to 1-sided training and testing using the common evaluation condition, restricted to English dialogue as detailed in the NIST evaluation plan [13]. The performance measures used for system evaluation were the equal error rate (EER) and minimum decision cost function (DCF)

Further experiments involving score-normalisation were conducted on the systems to aid in the comparison of the GMM and SVM domains [6]. A set of 55 male and 87 female T-Norm models were trained to estimate the score normalisation parameters.

4.1 GMM-UBM System

As a point of reference, a baseline GMM-UBM system was implemented. The system uses MAP adaptation with an adaptation factor of 8 and feature-warped MFCC features with appended delta coefficients [7]. Throughout the trials, 512 GMM mixture components were used. Gender dependent UBMs were trained using a diverse selection of 1818 utterances from both Mixer and Switchboard 2 corpora.

The GMM-UBM system employing the session variability modelling technique presented in [5] used gender dependent transform matrices $\boldsymbol{U}$ with a session subspace dimension of $R_s = 50$ trained from the same data as was used to train the UBMs.

4.2 GMM Supervector SVM System

The training of SVM speaker models required the production of several sets of utterance supervectors. A GMM mean supervector was produced to represent each *(1)* utterance in the background data set, *(2)* training utterance and *(3)* testing utterance. The background dataset consisted of all utterances used to train the UBMs.

The difference between the standard SVM and the session variability modelling SVM system is the method used to train the GMMs to represent each utterance prior to extraction of the supervectors. The baseline SVM system used standard MAP adapted GMMs to represent each utterance while the session SVM system employed session variability modelling training. In the latter system, session variability modelling was applied to the GMM of each utterance including those in the background data set.

For both systems, the supervectors were used to train one-sided SVM speaker models using LIBSVM [14]. A single supervector was used to represent the target training utterance while non-target training utterances were represented by the gender dependent background data set.

The SVM employed a linear-based kernel using background data scaling as detailed in (3).

5 Results

A comparison of performance between different system configurations is shown in Figure 1 with resulting EER and minimum DCF points detailed in Table 1. Results of systems including score normalisation are also detailed in this table.

These results show that a distinct performance gain can be achieved in discriminative classifiers when robust modelling techniques are applied during generative model training. This is evident by the observed performance variation between the two discriminative classifiers. The minimum DCF of the SVM system was reduced from .0258 to .0185 when session variability modelling was applied; a 28% relative improvement. In terms of EER, the session SVM system has a gain of 13% over the reference SVM configuration.

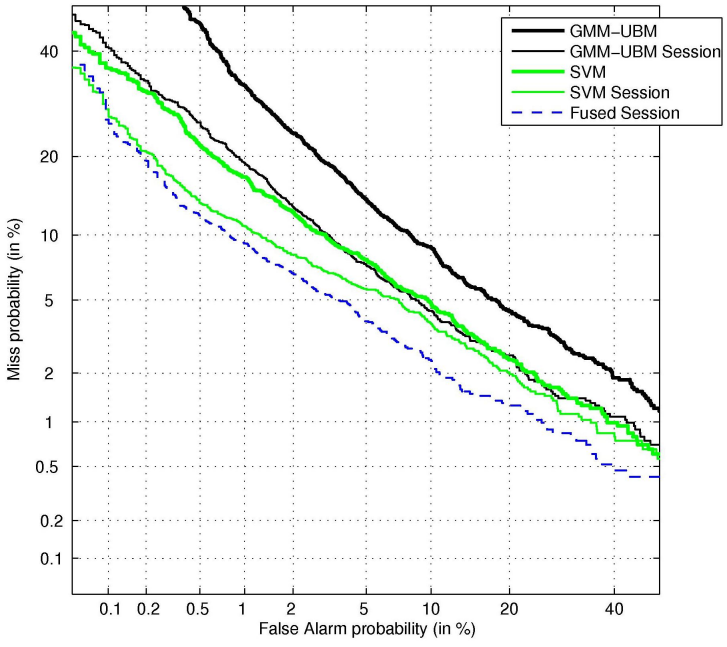

Fig. 1. DET plot for the 1-side condition comparing GMM-UBM and GMM mean supervector SVM systems, with and without session variability modelling.

A comparison between the reference GMM-UBM and SVM systems shows the SVM configuration having a gain of 38% in minimum DCF and 30% in EER over the GMM-UBM. Similarly, an improvement of 35% and 10% in minimum DCF and EER respectively is found between the two session configurations.

A significant improvement is shown through the GMM supervector SVM classification over the baseline GMM-UBM configuration which reflects the findings in [12]. Noteworthy is the performance of the reference SVM system being similar to that of the session GMM-UBM system throughout the mid to high false alarm range.

Table 1. Minimum DCF and EER results for 1-side condition for GMM-UBM and GMM mean supervector SVM systems, including T-Norm results.

System	**Standard**		**T-Norm**	
	EER	Min. DCF	EER	Min. DCF
Reference GMM-UBM	9.15%	.0418	9.95%	.0392
Session GMM-UBM	6.23%	.0286	5.58%	.0239
Reference SVM	6.38%	.0258	6.15%	.0240
Session SVM	5.58%	.0185	5.26%	.0189
Session Fused	4.41%	.0168	4.74%	.0160

Table 1 shows that a significant advantage was found through the application of T-Norm to the session GMM-UBM configuration, supporting previous results indicating that the session GMM-UBM system responds particularly well to score normalisation [5]. Conversely, the session SVM system showed little change through the normalisation technique while the reference GMM-UBM and SVM configurations both showed similar, moderate improvements when applying score-normalisation. The modest improvements due to T-Norm for the session SVM system suggests that this system may produce scores that are less prone to output score variations across different test utterances.

The scores from both the session GMM-UBM and the session SVM system were linearly fused to minimise the mean-squared-error. The DET plot demonstrates that performance is further boosted through this process. The fused system gave a relative improvement of 9% in minimum DCF and 21% in EER over the session SVM configuration. This result indicates that complementary information is found between the two systems despite session variability modelling being incorporated in both. Applying T-Norm to this fused system provided mixed results.

Future work will investigate further score normalisation methods for the GMM mean supervector SVM system using session variability modelling. A comparison between Campbell's method of nuisance attribute projection (modelling session variability in the SVM kernel) with GMM supervectors [12] and the work presented in this paper would also be of interest.

6 Conclusions

This paper has demonstrated that employing robust modelling techniques during GMM training improves the performance of a GMM mean supervector SVM speaker verification system. This is of interest due to the fundamental differences between the two classification systems; GMM's based on distribution estimation versus margin maximisation in SVM classification.

Applying session variability modelling during the training of the GMM mean supervectors for SVM classification showed significant performance gains when evaluated using the NIST 2005 SRE corpus and was superior to the session GMM-UBM configuration. Fusion of the session GMM-UBM and session SVM systems displayed performance above either configuration on its own.

Acknowledgments This research was supported by the Australian Research Council (ARC) Discovery Grant Project ID: DP0557387.

References

1. Reynolds, D., Quatieri, T., Dunn, R.: Speaker Verification Using Adapted Gaussian Mixture Models. Digital Signal Processing **10**(1) (2000) 19–41
2. Gauvain, J., Lee, C.: Maximum a posteriori estimation for multivariate Gaussian mixture observations of Markov chains. IEEE Transactions on Speech and Audio Processing **2**(2) (1994) 291–298
3. Przybocki, M., Martin, A.: NIST Speaker Recognition Evaluation Chronicles. Odyssey Workshop (2004)
4. Kenny, P., Dumouchel, P.: Experiments in speaker verification using factor analysis likelihood ratios. Odyssey: The Speaker and Language Recognition Workshop (2004) 219–226
5. Vogt, R., Sridharan, S.: Experiments in Session Variability Modelling for Speaker Verification. IEEE International Conference on Acoustics, Speech and Signal Processing **1** (May 2006) 897–900
6. Auckenthaler, R., Carey, M., Lloyd-Thomas, H.: Score normalization for text-independent speaker verification systems. Digital Signal Processing **10**(1) (2000) 42–54
7. Pelecanos, J., Sridharan, S.: Feature warping for robust speaker verification. Proc. Speaker Odyssey **2001** (2001)
8. Campbell, W., Sturim, D., Reynolds, D.: Support vector machines using GMM supervectors for speaker verification. Signal Processing Letters **13**(5) (May 2006) 308–311
9. Vogt, R.: Automatic Speaker Recognition Under Adverse Conditions. PhD thesis, Queensland University of Technology, Brisbane, Queensland (2006)
10. Reynolds, D.: Comparison of background normalization methods for text-independent speaker verification. Proc. Eurospeech **97** (1997)
11. Cortes, C., Vapnik, V.: Support-vector networks. Machine Learning **20**(3) (1995) 273–297
12. Campbell, W., Sturim, D., Reynolds, D., Solomonoff, A.: SVM Based Speaker Verification using a GMM Supervector Kernel and NAP Variability Compensation. IEEE International Conference on Acoustics, Speech and Signal Processing **1** (May 2006) 97–100

13. The NIST 2006 Speaker Recognition Evaluation Plan. (2006) Available at http://www.nist.gov/speech/tests/spk/2004/SRE-04_evalplan-v1a.pdf.
14. Chang, C.C., Lin, C.J.: LIBSVM: a library for support vector machines. (2001) Software available at http://www.csie.ntu.edu.tw/ cjlin/libsvm.